IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alyssa McDivitt
June 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- The SpaceX data was collected through using the SpaceX API and through webscraping of the spaceX wikipedia page. The data was then wrangled and Exploratory Data Analysis was performed using SQL queries and data visualisation (plotting seaborne graphs). Interactive visual analytics were created using Folium and Plotly Dash and helped collect insights on the dataset. Finally, predictive classification models were tested. Different classification models (Support Vector Machines (SVM), Classification Trees, and Logistic Regression) were built using the training set and tested using the test set.

- The best performing methods are logistic regression, support vector machine and k-nearest neighbours. They all have the same confusion matrix and the same accuracy score of 0.8333333333333334.

Introduction

- In this capstone project, our objective is to predict whether the Falcon 9 first stage will land successfully. The advertised price of a Falcon 9 rocket launch (on Space X's website) is 62 million US dollars each whereas other providers advertise costs upwards of 165 million US dollars. The difference in price can be explained by the fact SpaceX reuses the first stage of their rocket launch. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch. This project involves data science techniques such as data analysis, data visualization, and machine learning by using Python.

- The project is split into 4 parts, where the data is first collected through webscraping and focus is given on understanding the dataset. The data is then wrangled in Python. Exploratory Data Analysis is performed using SQL and using Pandas and Matplotlib in Python). Interactive visual analytics were then created using FoliumBuild and an interactive dashboard was created in Plotly Dash. Finally, predictive classification models were tested and validated in python. The data driven insights will be presented in this report.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX launch data was collected with the SpaceX REST API (Application programming interface). A python code was then written to convert a JSON file and manipulate data in a Pandas data frame.
  - Data was also sourced from the SpaceX Wikipedia page via webscraping, using the BeautifulSoup python package.
- Perform data wrangling
  - The data was parsed into a pandas dataframe and cleaned up to contain only information useful to this study. Data was wrangled by creating calculations and determining useful training labels.

Methodology

Executive Summary

- Exploratory data analysis (EDA) using visualization and SQL:
  - The dataset was queried and explored using SQL methods and seaborn visual plots.
- Interactive visual analytics using Folium and Plotly Dash:
  - An interactive folium map was created to geographically query and visualise the launch data.
  - An interactive plotly dash app was created to interactively slice and plot the data (scatter and pie charts).
- Predictive analysis using classification models:
  - The dataset was standardised then split into training and testing datasets.
  - Different classification models (Support Vector Machines (SVM), Classification Trees, and Logistic Regression) were built using the training set. The models were tuned by performing a grid search to find the hyperparameters that allow a given algorithm to perform best. The models were then evaluated by using confusion matrices and calculating accuracies on the test set. The models that performed best had the highest accuracy values.

Data Collection Summary

- **SpaceX API**: SpaceX launch data was collected with the SpaceX REST API. A GET request was used to request and parse the SpaceX launch data. The response content Json was decoded using .json() and turned it into a Pandas dataframe using .json_normalize(). The SpaceX API allows us to gather data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

- **Web scraping**: Falcon 9 launch records saved in HTML tables available on the SpaceX Wikipedia page were webscraped using the BeautifulSoup python library. The table was then parsed and converted into a pandas data frame.

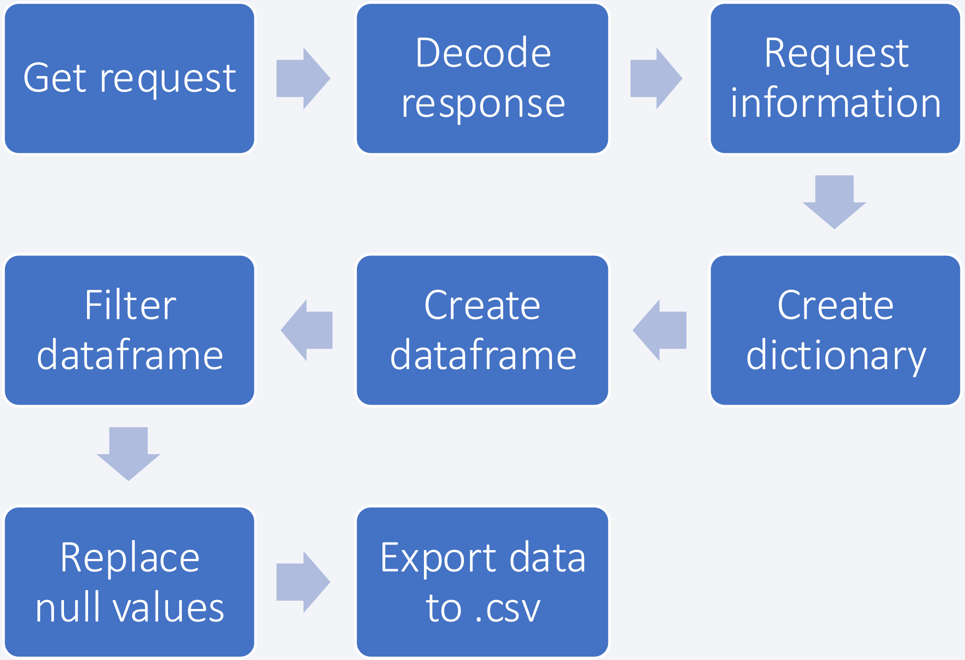
Data Collection – SpaceX API

Steps required for data collection through SpaceX API:

- Get request: request data from SpaceX API
- Decode response: using json() function
- Request information: request further information aboutusing columns rocket, payloads, launchpad, and cores. using API
- Create dictionary: combine the collected columns into a dictionary.
- Create dataframe: using .json_normalize()
- Filter dataframe: to contain only information about Flacon 9 launches
- Replace null values: replace missing payload mass with mean value
- Export data to a .csv

GitHub URL of the completed SpaceX API calls notebook:

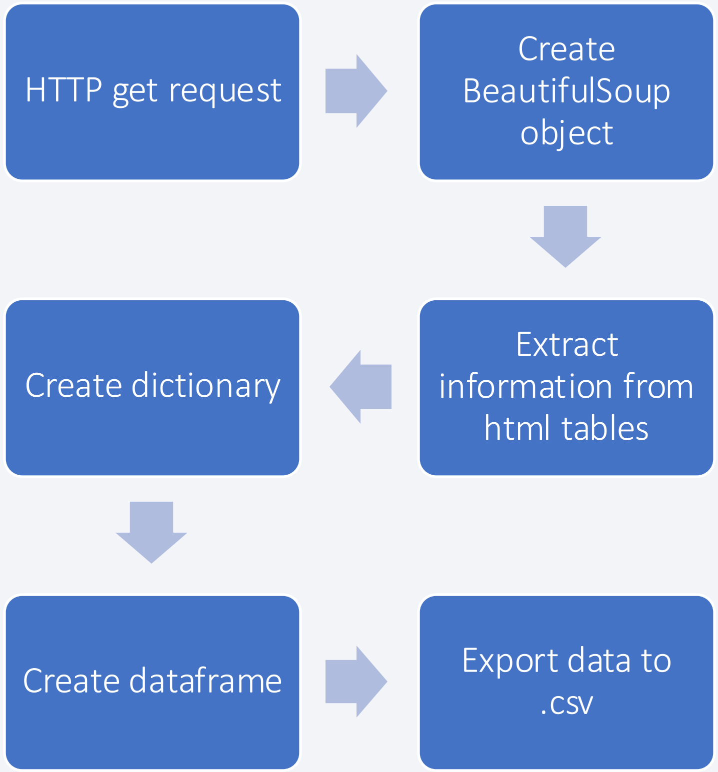
https://github.com/JustAGeologist/Data-Science-and-Machine-Learning-Capstone-Project/blob/8d91215fdc224ed58b2b41aa6030ec51e00689d3/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

Steps required for webscraping:

- HTTP GET method: to request the Falcon9 Launch HTML page, as an HTTP response.
- Create a BeautifulSoup object from a response text content
- Extract all column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables
- We will create an empty dictionary with keys from the extracted column names in the previous task.
- Later, this dictionary will be converted into a Pandas dataframe
- Present your web scraping process using key phrases and flowcharts
- GitHub URL: https://github.com/JustAGeologist/Data-Science-and-Machine-Learning-Capstone-Project/blob/f36928df0ed8158594e53ac3b3e2e5a45ab5f4b7/2-jupyter-labs-webscraping.ipynb

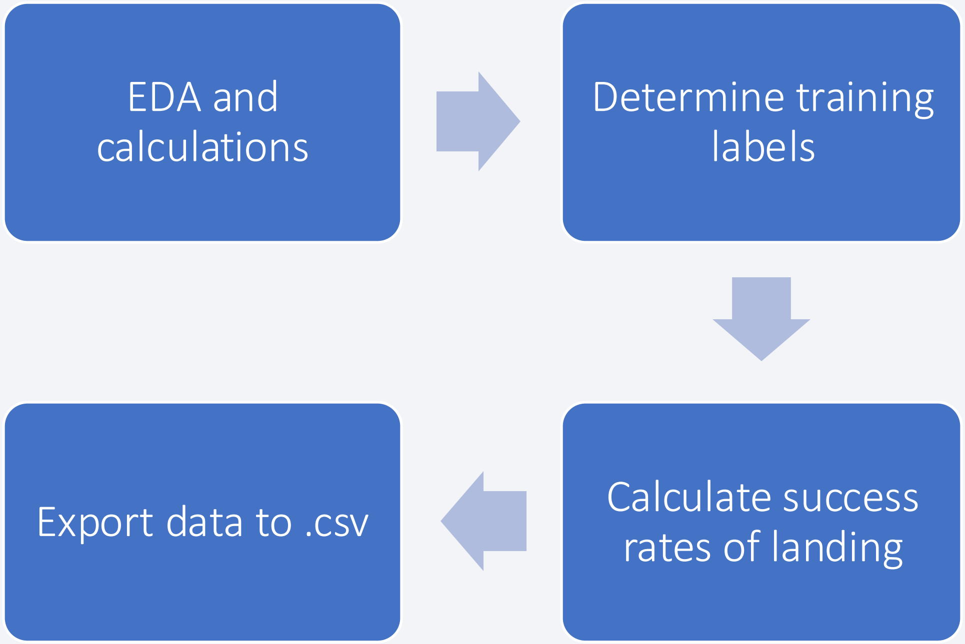


Data Wrangling

Data wrangling steps:

- Calculate the number of launches on each site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurence of mission outcome per orbit type
- Create a landing outcome label from Outcome column (convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful)
- Calculate success rate of landing
- Export data to .csv file

- GitHub URL: https://github.com/JustAGeologist/Data-Science-and-Machine-Learning-Capstone-Project/blob/62b6a2f3ab61aa2ce33029a7802603fcfd423347/3-labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb



EDA with Data Visualization

A variety of seaborne charts were plotted in this part of the study: categorical plots (catplots), barplots, scatter plots, and line plots)

- Categorical scatter plots: to visualize the relationship between Flight Number and Launch Site and the relationship between Payload and Launch Site
- Bar plot: to visualize the relationship between success rate of each orbit type
- Categorical scatter plot: to visualize the relationship between Payload and Orbit type
- Line plot: to visualize the launch success yearly trend. It represents a trend over time.

- GitHub URL: https://github.com/JustAGeologist/Data-Science-and-Machine-Learning-Capstone-Project/blob/ccc36bd5ba3172b10bf8875dbdada9275ba2d762/5-jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

Steps required:

- Load the SQL extension and connect to the @my_data1.db database
- Remove blank rows from table
- Display the names of the unique launch sites in the space mission using the DISTINCT function and the total payload mass carried by boosters launched by NASA (CRS) by using the SUM function. The average payload mass carried by booster version F9 v1.1 by was also displayed using the AVG function.
- List the date where the succesful landing outcome in drone ship was achieved.using the min() function on date
- List the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass using a subquery
- List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017. SQLLite does not support monthnames. So we used substr(Date,6,2) for month, substr(Date,9,2) for date, substr(Date,0,5),='2017' for year.
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- GitHub URL: https://github.com/JustAGeologist/Data-Science-and-Machine-Learning-Capstone-Project/blob/30a66ad3f1050a49fe8a4518d239bfa5bcdc1d6d/4-jupyter-labs-eda-sql-edx_sqllite.ipynb

Build an Interactive Map with Folium

Step required to create Folium Map output:

- Mark all launch sites with a circle and label on a map to make them easily visible. Depending on the proximity of launch sites, markers may be clustered.
- Mark the success/failed launches for each site on the map using colour coded circle markers.
- Calculate the distances between a launch site to its proximities. Lines were drawn and distances were maked on the map to easily assess the distance between launch sites and proximal locations (cities, coasts, highways, railways) to determine some geographical patterns about launch sites.
- The map is interactive and allows the user to zoom in and out as well as get coordinates from mouse locations.
- GitHub URL: https://github.com/JustAGeologist/Data-Science-and-Machine-Learning-Capstone-Project/blob/0957acd5bfa16c9480838581330 3ad82e1645dfc/6-lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

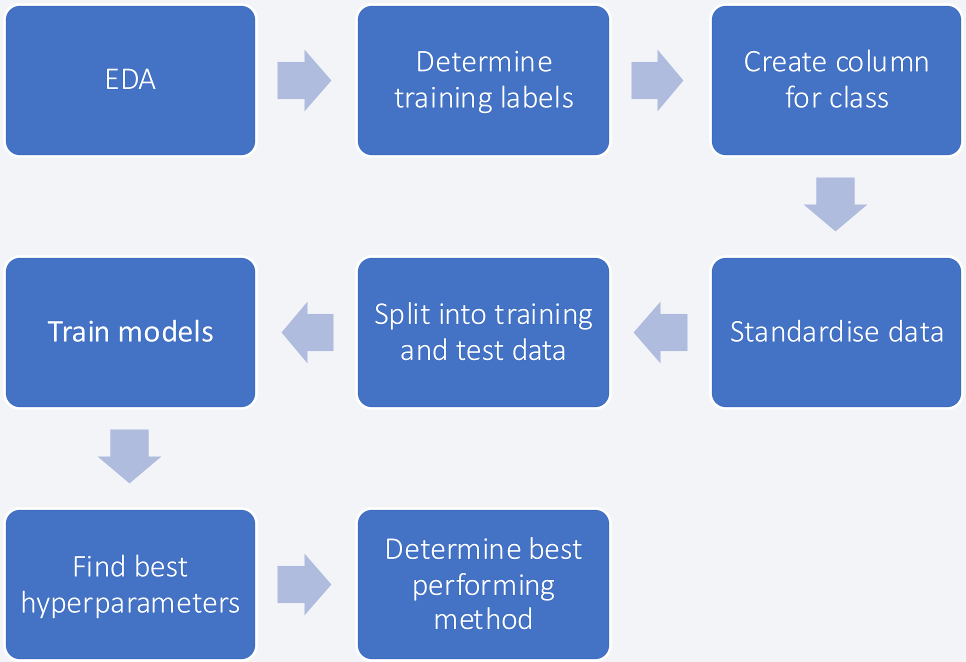
- An interactive dashboard was created using Plotly Dash
- An interactive dropdown list allows the dashboard user to select a specific site or all sites, which refreshes the data in the pie chart below.
- A pie chart was plotted showing the total success launches by site
- A categorical scatter plot was created to illustrate the between payload mass and success rate for different booster versions. The payload mass range can be selected by the user using an interactive scale bar. It refreshes the data in the scatter plot.

- GitHub URL: https://github.com/JustAGeologist/Data-Science-and-Machine-Learning-Capstone-Project/blob/3af1c88077f02006291bfdc3ea69bf1231a213f6/7-spacex_dash_app.py

Predictive Analysis (Classification)

Steps required:

- Dataset was standardised then split into training and testing sets
- Different classification models (Support Vector Machines (SVM), Classification Trees, and Logistic Regression) were built using the training set. The models were tuned by performing a grid search to find the hyperparameters that allow a given algorithm to perform best.
- The models were then evaluated by using confusion matrices and calculating accuracies on the test set. The models that performed best had the highest accuracy values.
- GitHub URL: https://github.com/JustAGeologist/Data-Science-and-Machine-Learning-Capstone-Project/blob/30dc6df849cb4b2a365195cb4ae9f234ee67a8f2/8-SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb



Results

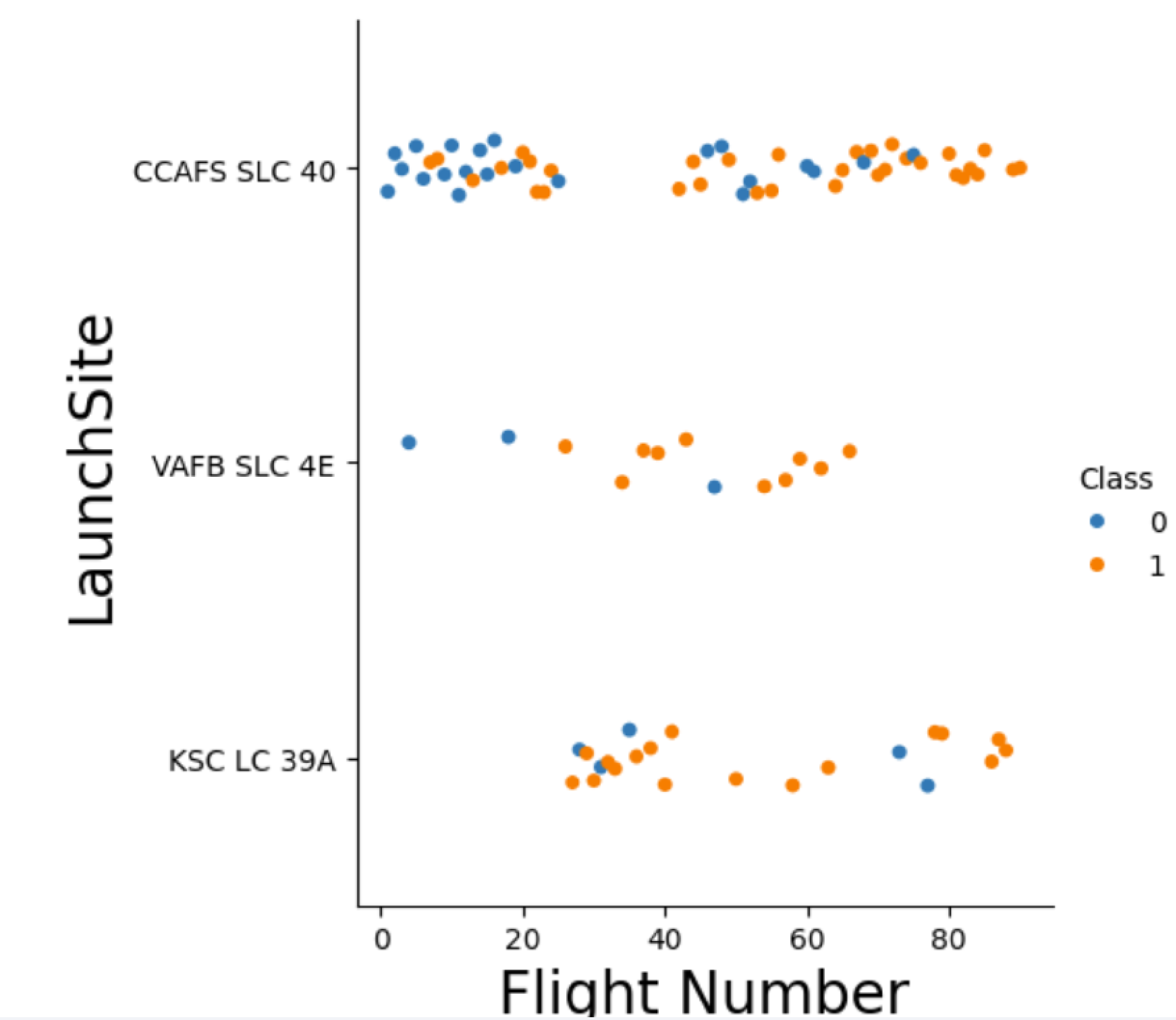
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

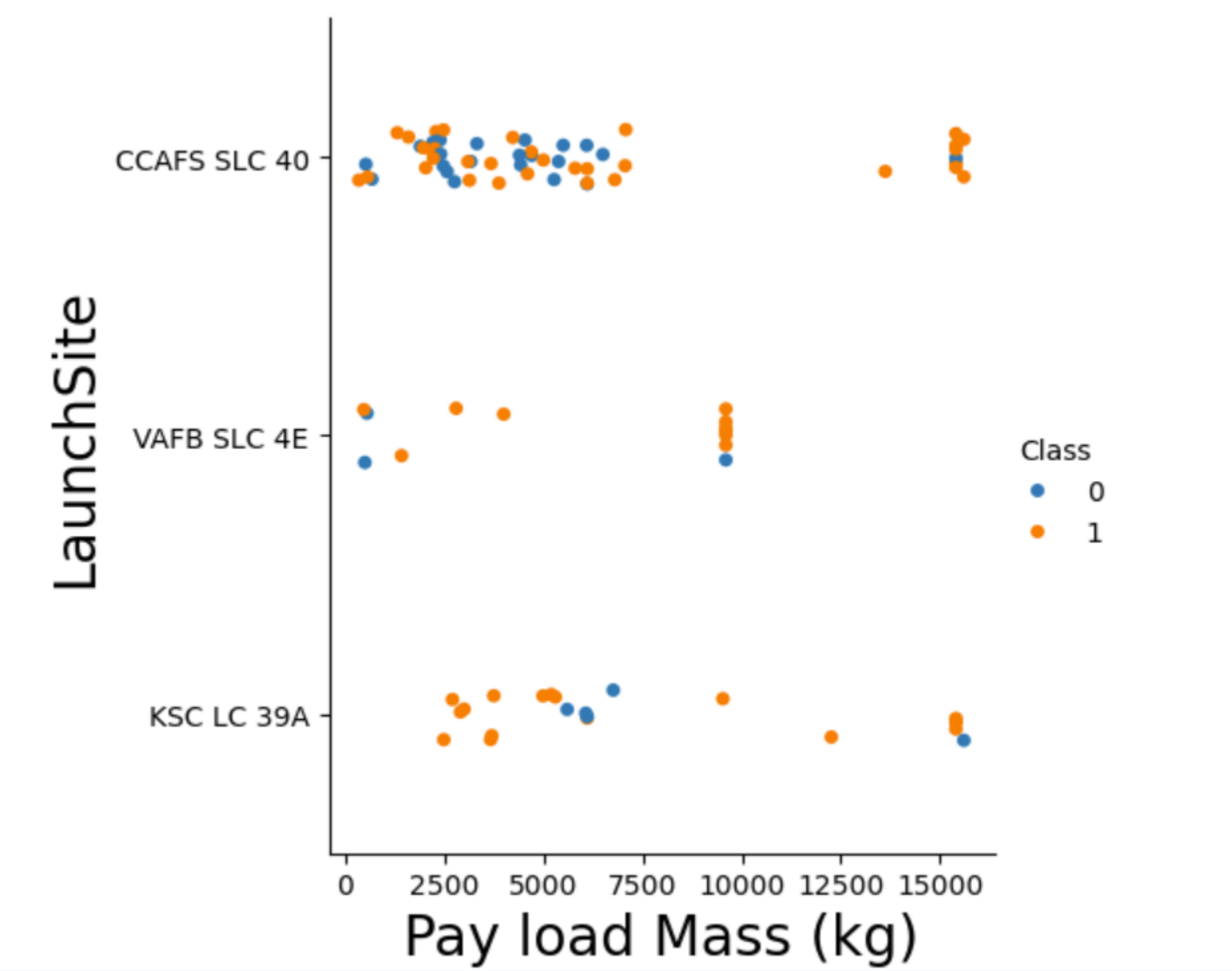
Flight Number vs. Launch Site

- This categorical scatter plot suggests that launch site CCAFS SLC 40 was most used (highest number of points).
- With an increasing flight number, the class evolves from 0 (failed launches, in blue) to a higher frequency of 1 values (successful launches, in yellow)



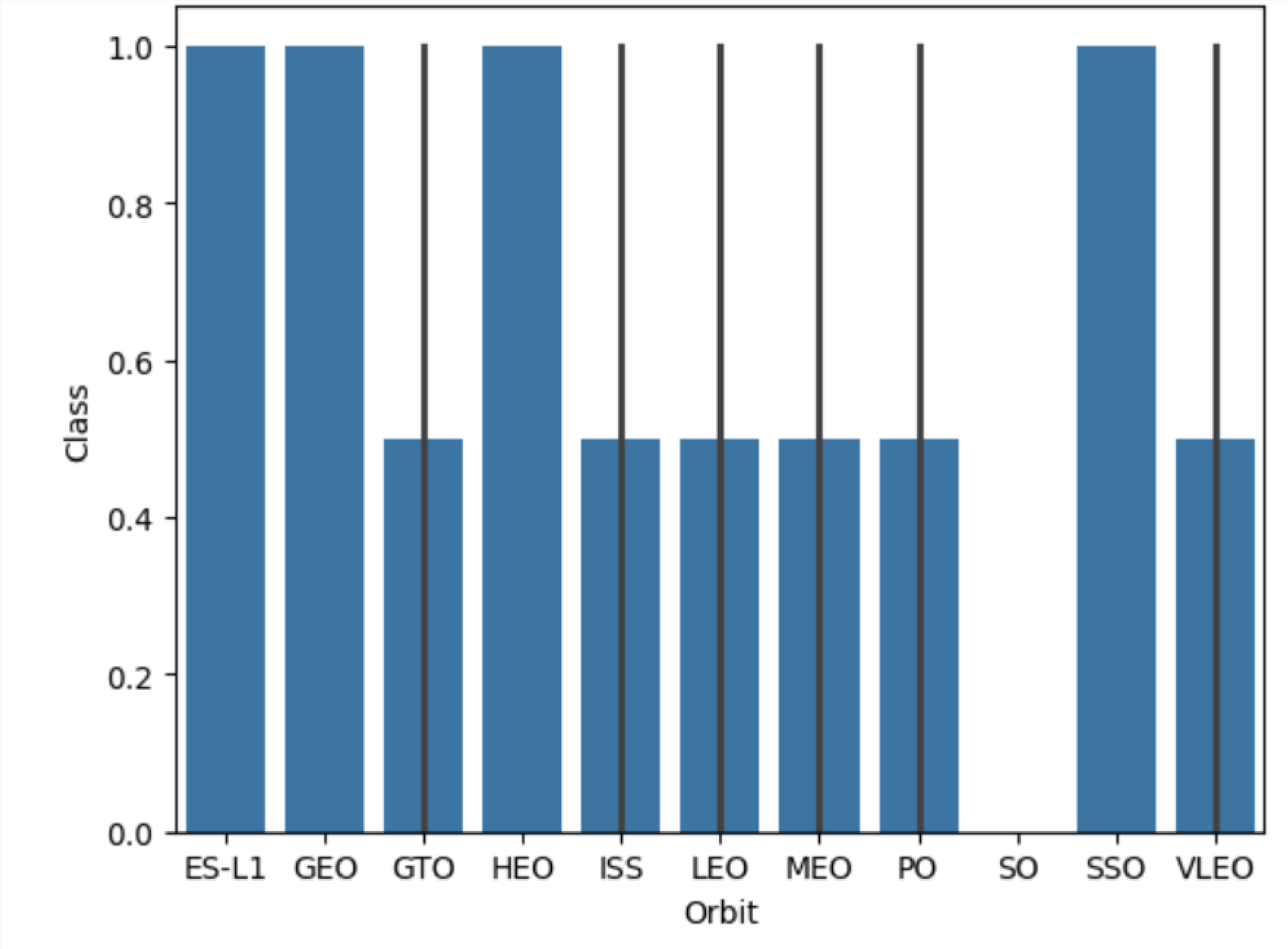
Payload vs. Launch Site

- The scatter plot of Payload vs. Launch Site shows that with higher payload masses (7500 kg), the launches have a higher chance of being successful (class 1).
- Launch site VAFB SLC 4E has not launched a payload greater than 10 000 kg.



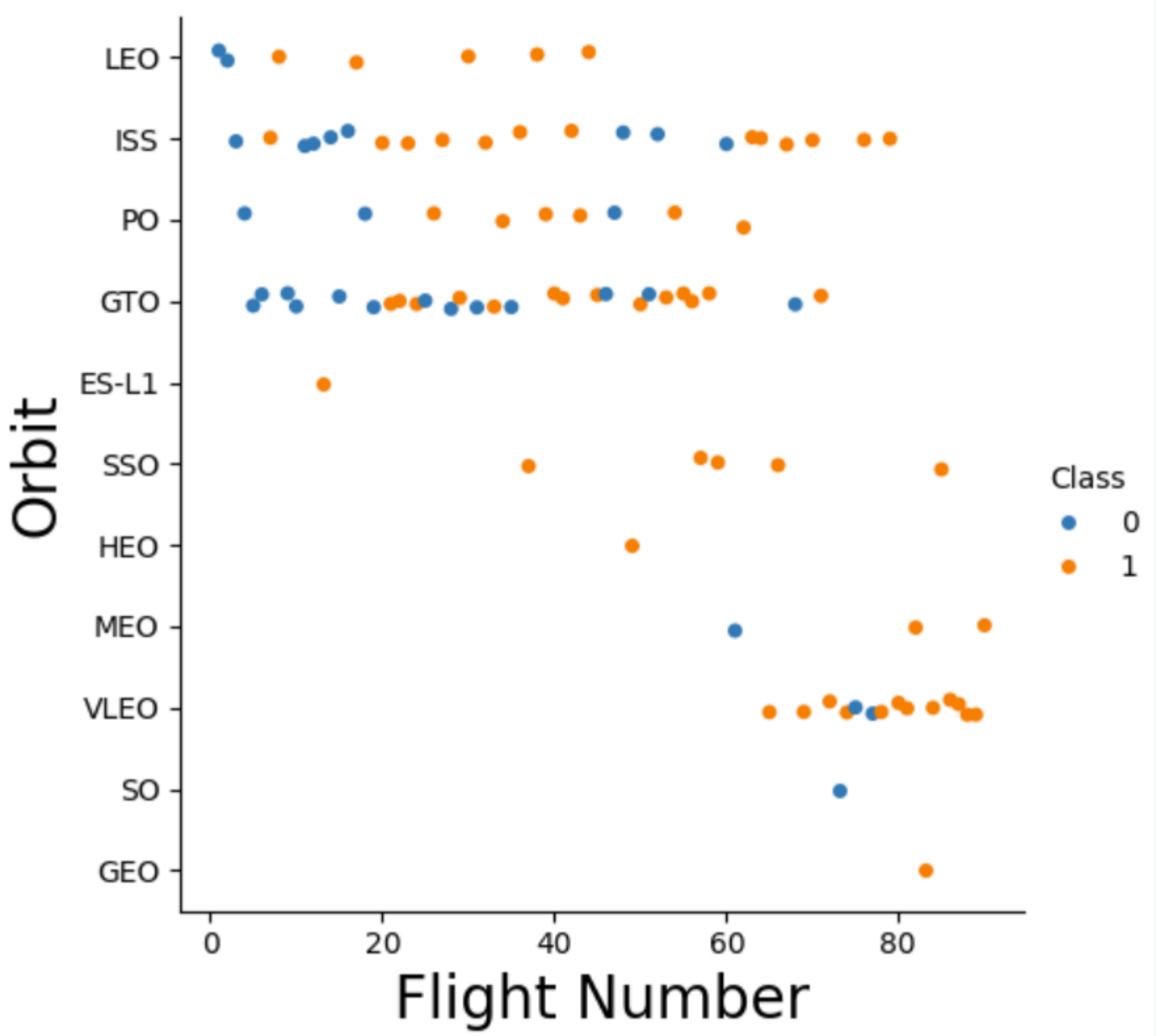
Success Rate vs. Orbit Type

- The bar chart for the success rate of each orbit type shows that orbit types ES-L1, GEO, HEO and SSO have a 100% success rate (Class 1) whereas orbit types GTO, ISS, LEO, MEO, PO and VLEO have success rate of 50%.
- Orbit type SO has a 0% success rate.



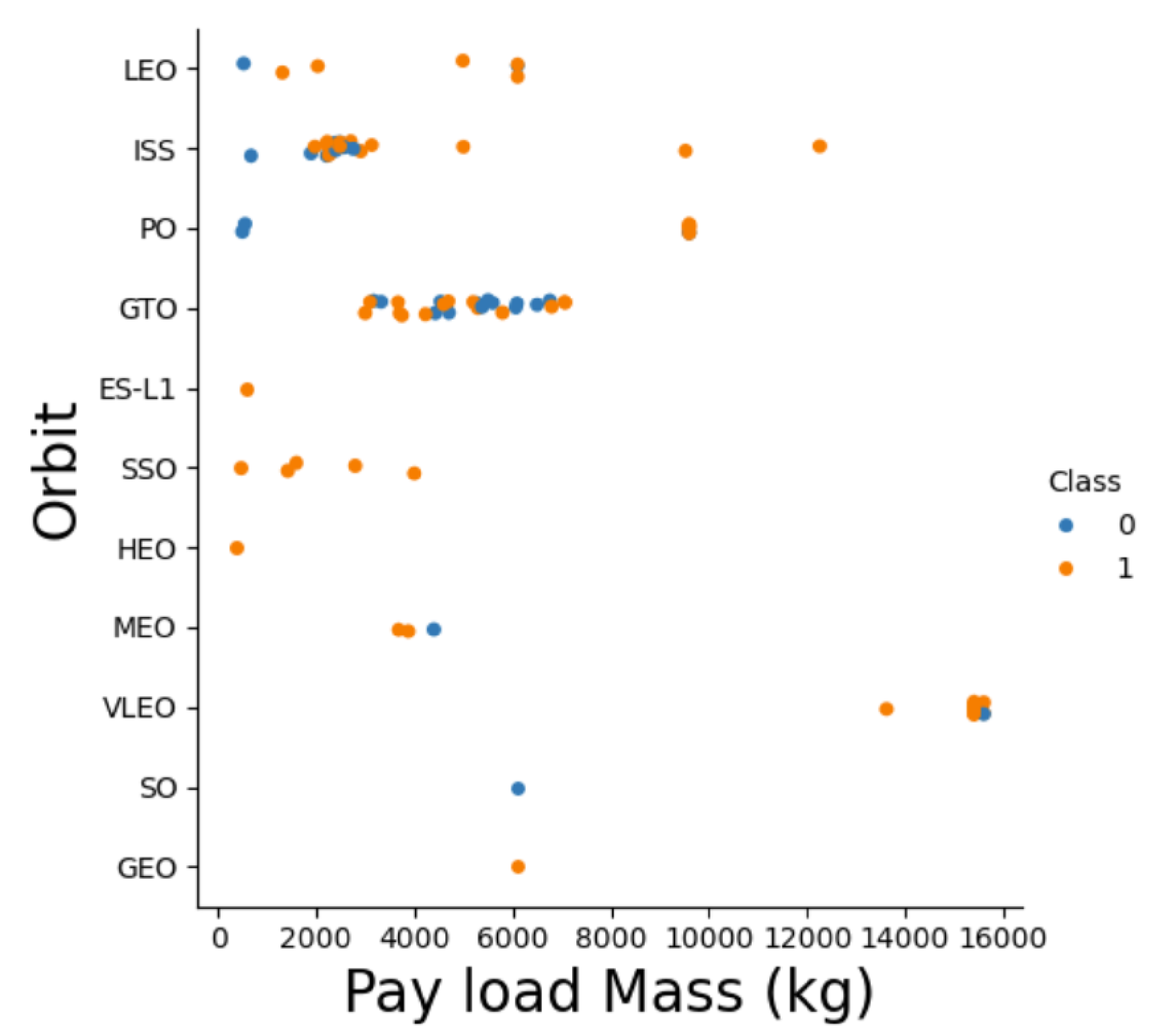
Flight Number vs. Orbit Type

- A number of observations can be drawn from the scatter point of Flight number vs. Orbit type:
  - Although Orbits ES-L1, SSO, and GEO have very low number of launches, they have a 100% success rate (class 1)
  - Orbits ISSandPO have a higher number of launches and with an increasing flight number, their success rate has increased
  - Although orbit GTO has a high number of launches, the results are mixed (moderate success rate)
  - Only one launch was performed in Orbit SO and it failed (class 0).



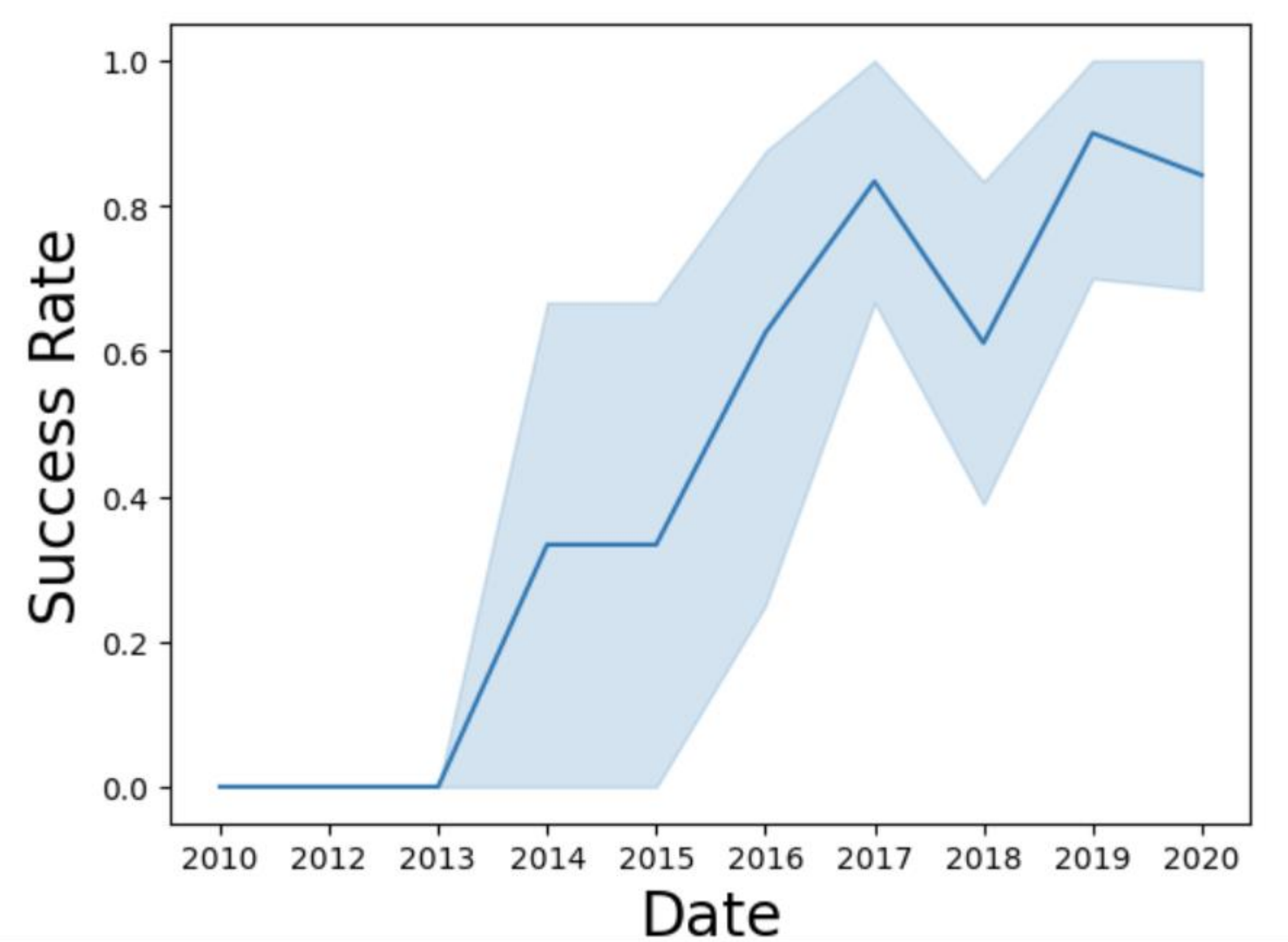
Payload vs. Orbit Type

- A number of observations can be made from the scatter point of payload vs. orbit type:
    - The higher the payload mass, the lower the number of launches.
    - The launches performed with higher payload masses (>8000 kg) are generally successful.
    - The launches performed in orbit SSO, HEO and ES-L1 have been carried out with low payload masses and have been 100% successful.
    - The launches in orbit GTO show mixed results (successes and fails regardless of the difference in payload mass)
    - Launches in orbits LEO, PO and ISS show that launches with higher payload masses are more successful than with lighter payload masses.



Launch Success Yearly Trend

- The line chart of yearly average success rate clearly shows an increase in success rate as the years go by, with the exception for 2018 where a significant drop in success rate occurred.



All Launch Site Names

- The names of the unique launch sites were identified using the DISTINCT function in the SQL query pasted below:

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

- The distinct launch site names are presented in the table below:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'KSC'

- The query below was used to find 5 records where launch sites' names start with `KSC`. The like function allows to find names that start with KSC. The LIMIT function limits the selection to the first 5 names.

```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'KSC%' LIMIT 5;
```

- The outcome of the query is pasted below:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landin
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Succ
2017-03-16	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Suc
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Succ
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	

Total Payload Mass

- The total payload carried by boosters from NASA was calculated using the SUM function in SQL:

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
```

- The outcome of the query is pasted below:

SUM("PAYLOAD_MASS__KG_")
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 was calculated using the AVG function:

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTBL WHERE Booster_Version = 'F9 v1.1';
```

- The outcome of the query is pasted below:

AVG("PAYLOAD_MASS__KG_")
2928.4

First Successful drone ship Landing Date

- The dates of the first successful landing outcome on drone ship were extracted from the query below. The min function was used on the Date column:

```
%sql select min(Date) from SPACEXTBL where "Landing_Outcome" = "Success (drone ship)";
```

- .The date is:

min(Date)
2016-04-08

Successful Ground Pad Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on a ground pad and had payload mass greater than 4000 but less than 6000 were listed using the following query:

- ```
  %sql SELECT "Booster_Version" FROM SPACEXTBL WHERE
  "Landing_Outcome" = "Success (ground pad)"
  AND  "PAYLOAD_MASS__KG_"  > 4000 AND "PAYLOAD_MASS__KG_" <
  6000;
  ```

- The outcome of the query is summarised in the table below:

Booster_Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1
  ```

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes were calculated using this query: %sql SELECT Mission_Outcome, COUNT(Mission_Outcome) AS number_of_outcome FROM SPACEXTBL GROUP BY Mission_Outcome;

- Only one mission failed in flight whereas all other missions were successful.

Mission_Outcome	number_of_outcome
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass were found using the following query: %sql SELECT Booster_Version, "PAYLOAD_MASS__KG_" FROM SPACEXTBL WHERE "PAYLOAD_MASS__KG_" = (SELECT max("PAYLOAD_MASS__KG_") FROM SPACEXTBL);

- The outcome of the query is summarised in the table:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2017 Launch Records

- The month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017 records were extracted using the following query: %sql SELECT substr(Date, 6, 2) AS month, "Landing_Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTBL WHERE substr(Date,0,5)="2017" AND "Landing _Outcome" IN ("Success (ground pad)");
- Note: SQLLite does not support monthnames. So you need to use substr(Date,6,2) for month, substr(Date,9,2) for date, substr(Date,0,5),='2017' for year.

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- The following query was used to rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order: %sql SELECT * FROM SPACEXTBL WHERE "Landing_Outcome" = "Success (ground pad)" OR "Landing_Outcome" = "Failure (drone ship)" AND ("Date" BETWEEN '2010-06-04' AND '2017-03-20') ORDER BY "Date" DESC;
- The output of the query is summarised in the table in the next slide. Please note that the entire table could not fit in one page, please refer to GitHub link to see the full table: https://github.com/JustAGeologist/Data-Science-and-Machine-Learning-Capstone-Project/blob/30dc6df849cb4b2a365195cb4ae9f234ee67a8f2/4-jupyter-labs-eda-sql-edx_sqllite.ipynb

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- The output of the query is summarised in the table::

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Ou
2018-01-08	1:00:00	F9 B4 B1043.1	CCAFS SLC-40	Zuma	5000	LEO	Northrop Grumman	Success (p status u
2017-12-15	15:36:00	F9 FT B1035.2	CCAFS SLC-40	SpaceX CRS-13	2205	LEO (ISS)	NASA (CRS)	S
2017-09-07	14:00:00	F9 B4 B1040.1	KSC LC-39A	Boeing X-37B OTV-5	4990	LEO	U.S. Air Force	S
2017-08-14	16:31:00	F9 B4 B1039.1	KSC LC-39A	SpaceX CRS-12	3310	LEO (ISS)	NASA (CRS)	S
2017-06-03	21:07:00	F9 FT B1035.1	KSC LC-39A	SpaceX CRS-11	2708	LEO (ISS)	NASA (CRS)	S
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	S
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	S
2016-07-18	4:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	S
2016-06-15	14:29:00	F9 FT B1024	CCAFS LC-40	ABS-2A Eutelsat 117 West B	3600	GTO	ABS Eutelsat	S

Section 3

Launch Sites Proximities Analysis

Map of all launch sites

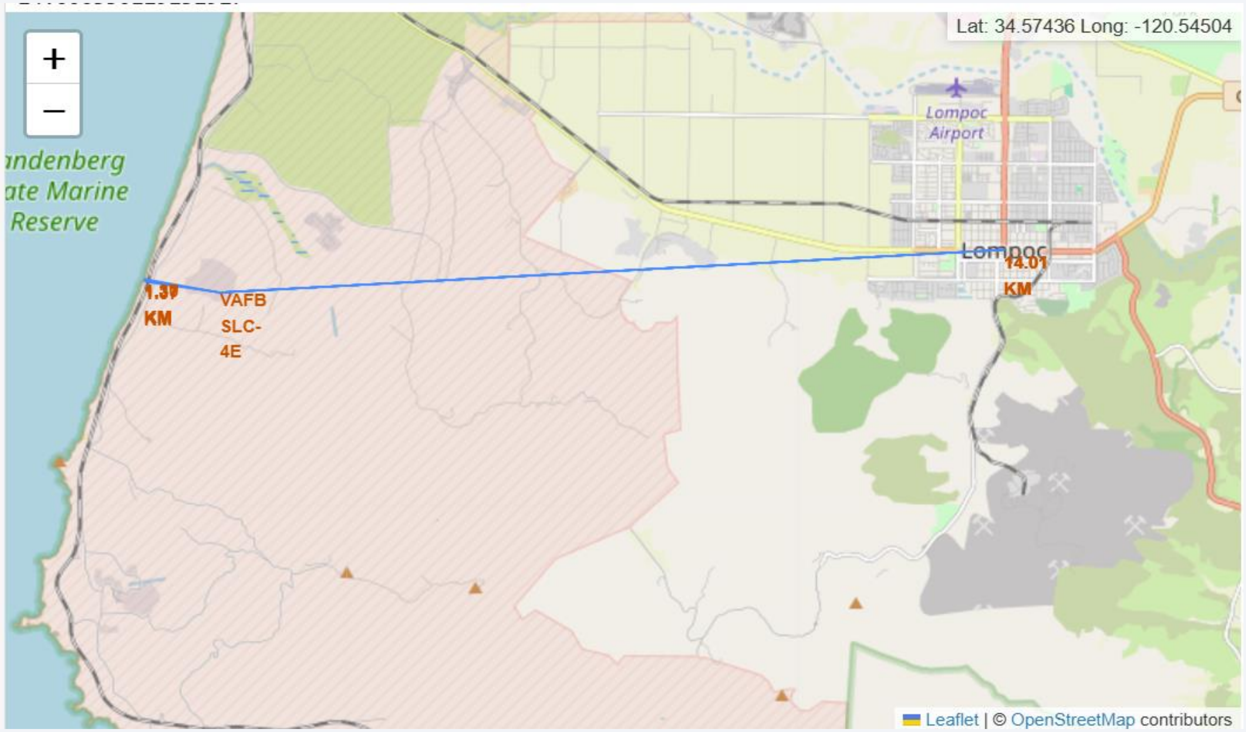
- The interactive folium map includes all launch sites' location markers.
- All launch sites seem to be in coastal areas the southern part of the US, close to the cancer tropic line.

<Folium Map Screenshot 2>

- Replace <Folium map screenshot 2> title with an appropriate title
- Explore the folium map and make a proper screenshot to show the color-labeled launch outcomes on the map
- Explain the important elements and findings on the screenshot

Map of proximity of launch site to the coast, city and highway

- Launch site VAFB SLC 4E is 1.39km from the highway and the coast, and 14.01km away from the nearest city
- The location is strategic for testing of rocket launches but also allows for supply of parts and workforce.

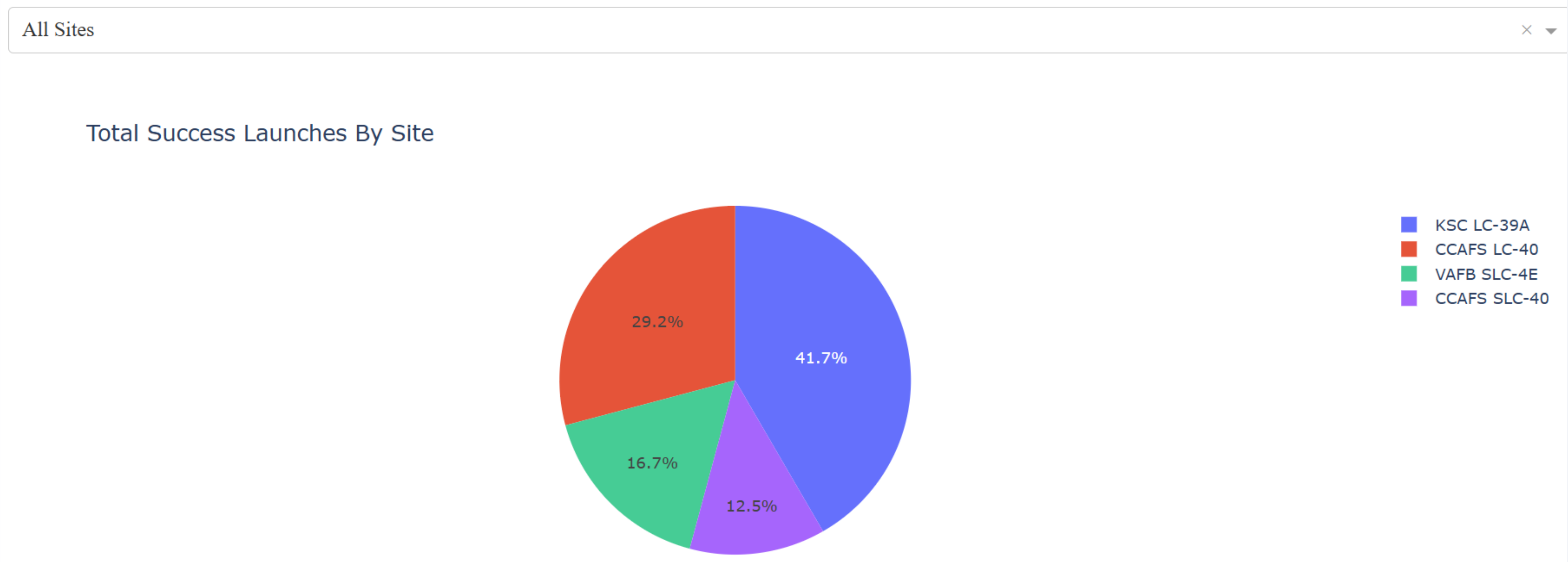


Section 4

Build a Dashboard with Plotly Dash

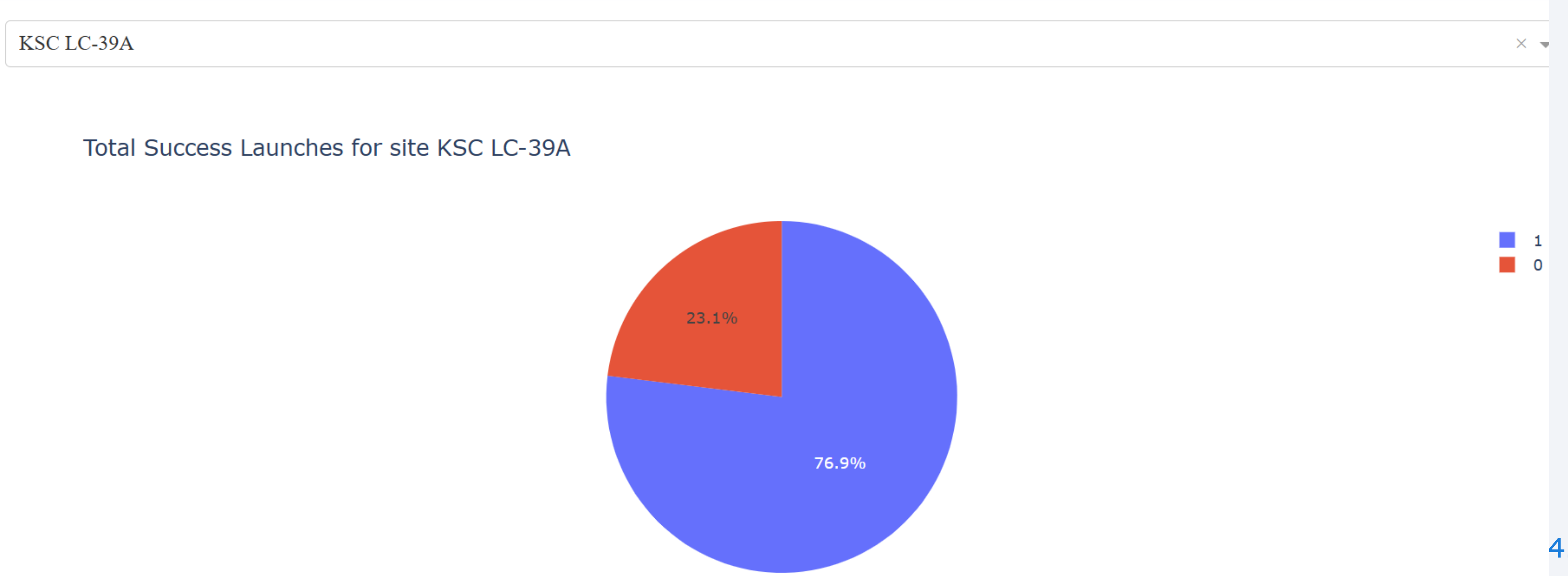
Piechart of launch succes counts for all sites

- Most (41.7%) successful launches have been launched from launch site KSC-LC-39A. The smallest proportion of successful launches were launched from CCAFS SLC 40.



Piechart for launch site KSC LC 39A

- KSC LC 39A is the launch site with highest launch success ratio (1= success and 0=fail)



Payload vs Launch Outcome scatter plot for all sites with payload range slider

- The payload range between 2000 kg and 4000 kg has the highest success rate (class=1 is success).
- FT booster versions have the highest success rate.

Section 5

Predictive Analysis (Classification)

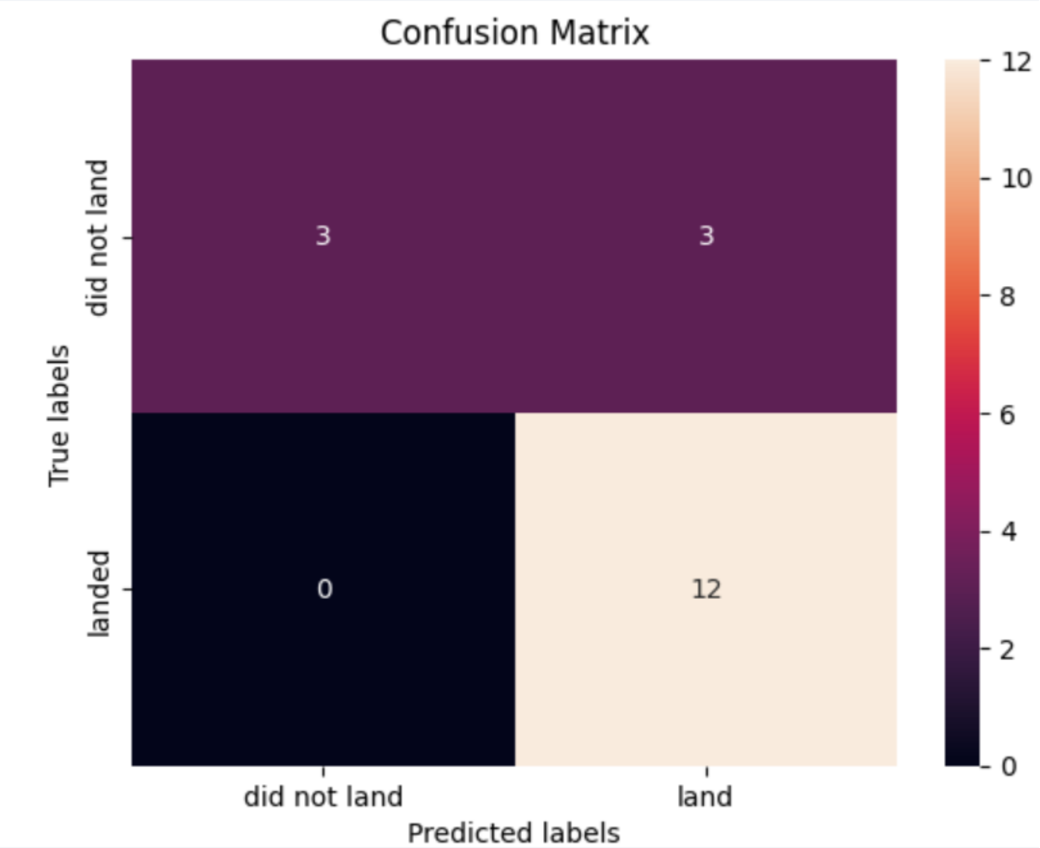
Classification Accuracy

- Visualize the built model accuracy for all built classification models, in a bar chart

- Find which model has the highest classification accuracy

Confusion Matrix

- The best performing methods are logistic regression, support vector machine and k-nearest neighbours. They all have the same confusion matrix and the same accuracy score of 0.8333333333333334.



Conclusions

- The best performing methods are logistic regression, support vector machine and k-nearest neighbours. They all have the same confusion matrix and the same accuracy score of 0.8333333333333334.

Thank you!